Intelligent Archives
A Conceptual Architecture Study

H. K. Ramapriyan, Gail McConaughy, Chris Lynnes, Steve Kempler, Ken McDonald

NASA Goddard Space Flight Center

Robert Harberts, Larry Roelofs,

Global Science and Technology, Inc.

IDU/AR Workshop

September 4-6, 2002

Ramapriyan@gsfc.nasa.gov

Goal

To create a next generation conceptual archive architecture supported by advanced technology that is able to:

- Increase data utilization by hosting and applying IDU technologies such as:
    - Information and knowledge extraction
    - Automated data object identification and classification
    - Intelligent user interfacing, and system management
    - Distributed computing and data storage
- Automate the transformation of data to information and knowledge allowing the user to focus on research/applications rather than data and data system manipulation
- Exploit new and emerging technologies as they become available
- Incorporate lessons learned from existing archives
- Accommodate new data intensive missions without redesign or restructuring

Technical Objectives

- Formulate concepts and architectures that support data archiving for NASA science research and applications in the 10 to 20 year time frame

- Focus on architectural strategies that will support intelligent processes and functions

- Identify and characterize science and applications scenarios that drive intelligent archive requirements

- Assess technologies and research that will be needed for the development of an intelligent archive

- Identify and characterize potential IDU research projects that will be needed to develop and create an intelligent archive

The Problem

- Data acquisition and accumulation rates tend to outpace the ability to access and analyze them

- The variety of data implies a heterogeneous and distributed set of data providers that serve a diverse, distributed community of users

- Unassisted human-based manipulation of vast quantities of archived data for discovery purposes is difficult and potentially costly

- If NASA is to migrate its technologies to operational agencies' decision support systems, it is necessary to demonstrate the feasibility of near-real-time utilization of vast quantities of data and the derived information and knowledge

- The types of data access and usage in future years are difficult to anticipate and will vary depending on the particular research or application environment, its supporting data sources, and its heritage system infrastructure

ESE supports 68 data centers (some of which at the same location), widely distributed geographically. Additional data centers, including NOAA's NCDC and Unidata, are networked through membership in the ESIP Federation.

Legend:
- ☆ Type 1 ESIPs (inc. DAACs)
- ◆ Type 2 ESIPs
- ○ Type 3 ESIPs
- ■ RESACs
- ⬢ Affiliated Res. Centers (ARCs)
- ▲ Directed
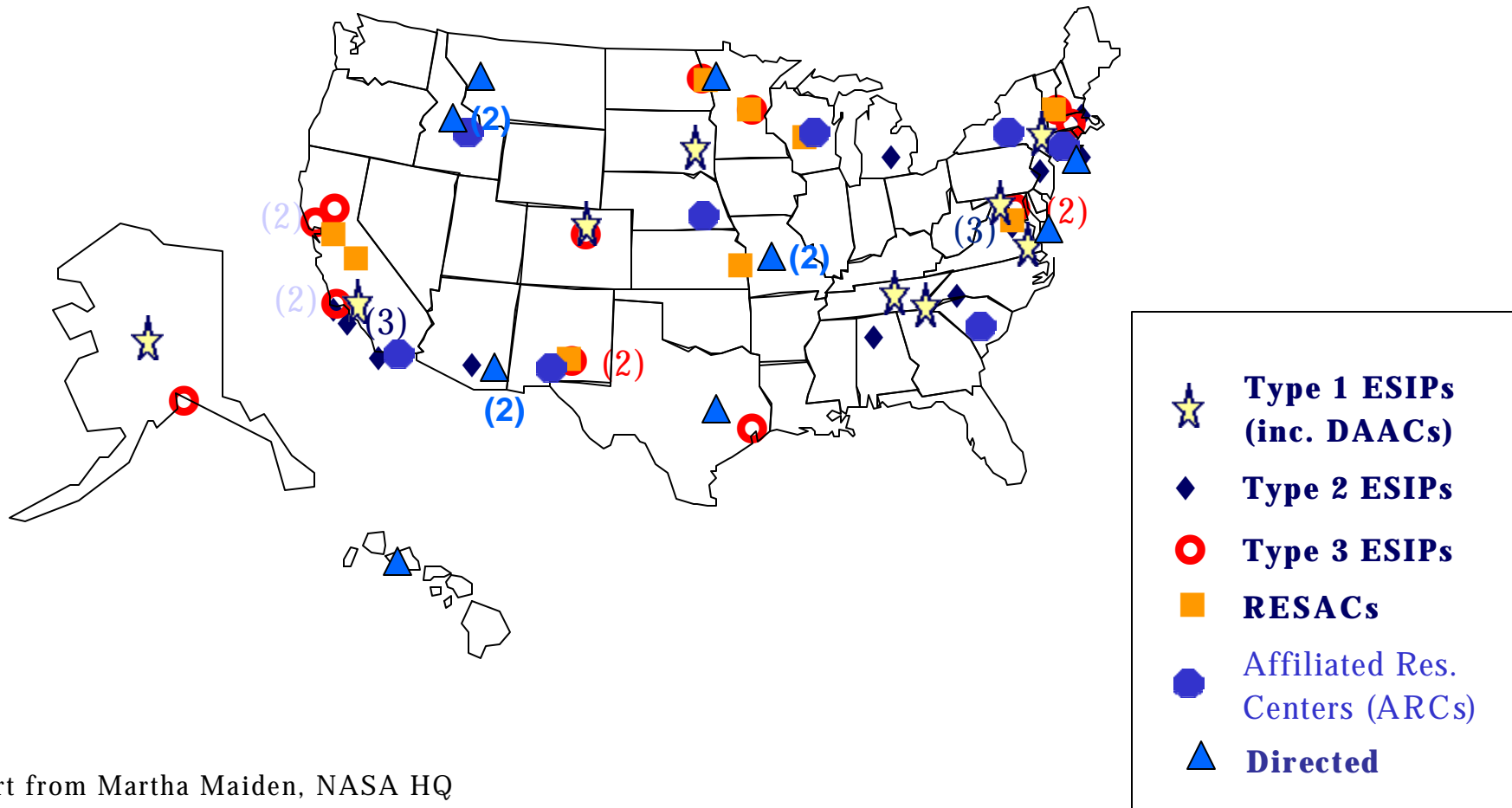
Chart from Martha Maiden, NASA HQ

Global Wireless Network

GPS

Other Pilots

Data production facility

Transportation Industry

Local wireless networks

Individual users
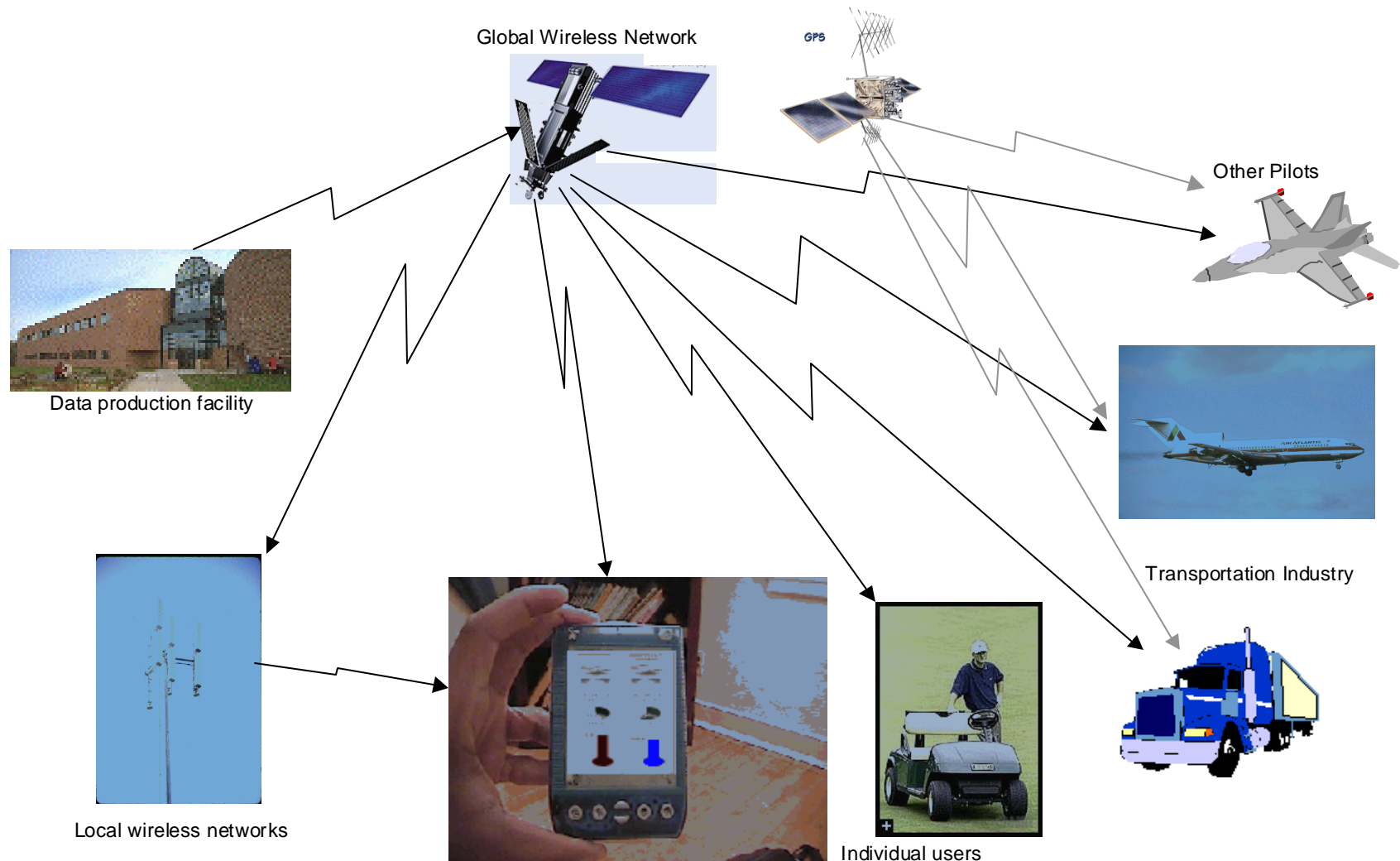
What Is An Intelligent Archive (IA)?

- An IA includes all items stored to support "end-to-end" research and applications scenarios
- Stored items include:
  - Data, information and knowledge
  - Software and processing needed to manage holdings and improve self-knowledge (e.g., data-mining to create robust content-based metadata)
  - Interfaces to algorithms and physical resources to support acquisition of data and their transformation into information and knowledge (could be invoked in push or pull mode)
- Architecture expected to be highly distributed so that it can easily adapt to include new elements as data and service providers
- Will have evolved functions beyond that of a traditional archive
- Will be based on and exploit technologies in the 10 to 20 year time range
- Will be highly adaptable so as to meet the evolving needs of science research and applications in terms of data, information and knowledge

Data, Information and Knowledge

Data: an assemblage of measurements and observations, particularly from sensors or instruments, with little or no interpretation applied

- Examples: *Scientific instrument measurements, market past performance*

Information: a summarization, abstraction or transformation of data into a more readily interpretable form
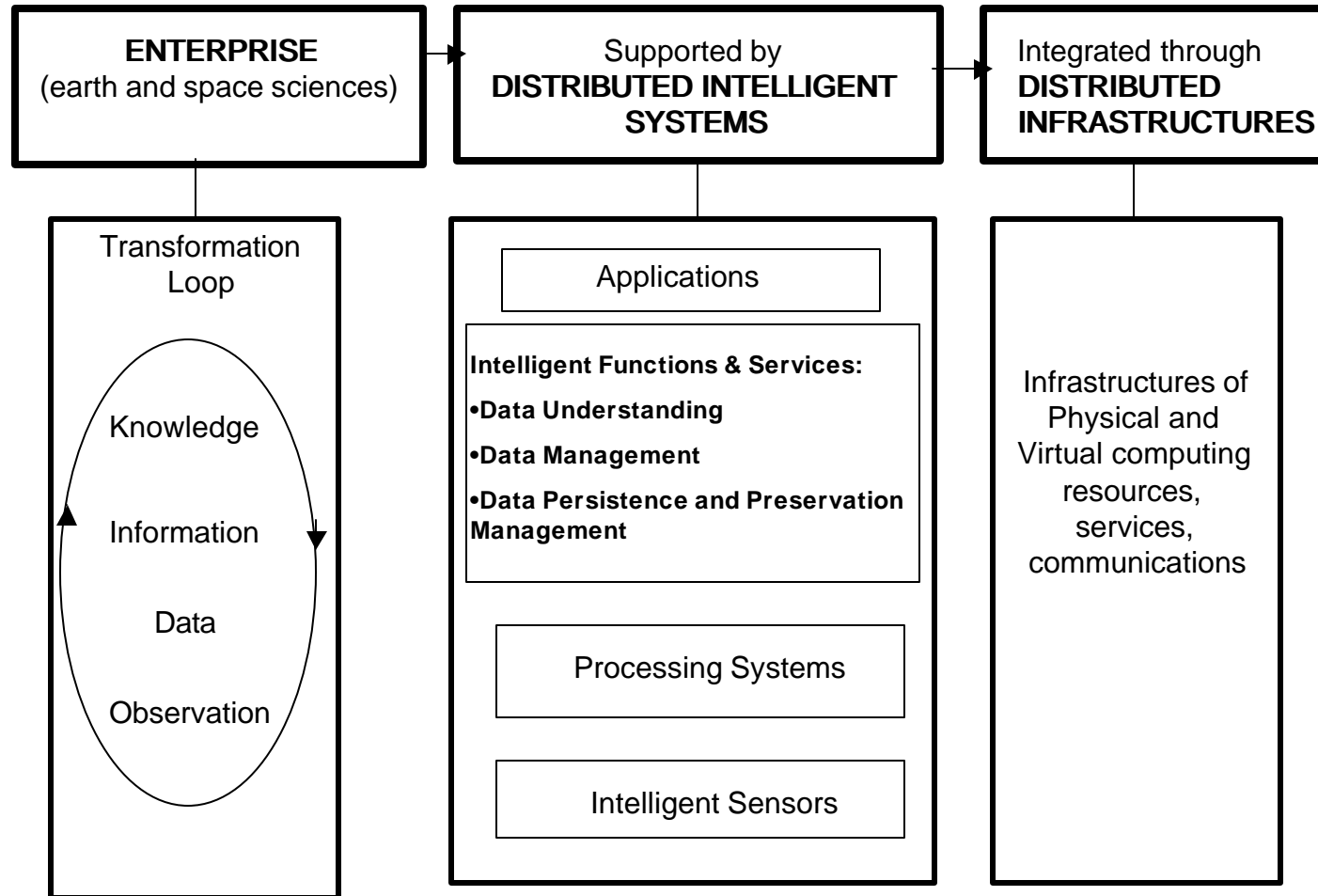
- Examples: *results after performing transformations by data mining, segmentation, classification, etc., such as a Landsat scene spatially indexed based on content, assigned a "class" value, fused with other data types, and subset for an application, for example a GIS.*

Knowledge: a summarization, abstraction or transformation of information that allows our understanding of the physical world

- Examples: *predictions from model forward runs, published papers, output of heuristics, or other techniques applied to information to answer a "what if" question such as "What will the accident rate be if an ice storm hits the Washington D.C. Beltway between Chevy Chase and the Potomac crossing at 7 a.m.?"*

Context - A Knowledge-Building System

ENTERPRISE
(earth and space sciences)

Supported by
DISTRIBUTED INTELLIGENT SYSTEMS

Integrated through
DISTRIBUTED INFRASTRUCTURES

Transformation Loop

Knowledge

Information

Data

Observation

Applications

Intelligent Functions & Services:
•**Data Understanding**
•**Data Management**
•**Data Persistence and Preservation Management**

Processing Systems

Intelligent Sensors

Infrastructures of Physical and Virtual computing resources, services, communications

Data Production System

Science Model System

Exploitation Tools

Cooperating Systems Interface

Register Lookup Broker

Intelligent Interim Archive

Intelligent Permanent Archive

Operations

Data Management

Added Value Content

Resource Infrastructure Interface

Self-Monitoring

Self-Adjusting

Self-Recovery

Performance Tuning

Cooperative Interface Status & Information Mgmt

Brokering
Store/Retrieve
Querying

Cataloging

Distributing

Registering

Change Formats

Mining

Characterization

Sub-setting

Fusion

Provisioning

Adapting

Sharing

- Storing and managing full representations of data, information, and knowledge
- Building intelligence about transformations on data, information, knowledge, and accompanying services involved in a scientific enterprise
- Performing self-analysis to enrich metadata that adds value to the archive's holdings
- Performing change detection to develop trending information
- Interacting as a cooperative node in a "web" of other systems to perform knowledge building (where knowledge building involves the transformations from data to information to knowledge) instead of just data pipelining
- Being aware of other nodes in the knowledge building system (participating in open systems interfaces and protocols for virtualization, and collaborative interoperability)
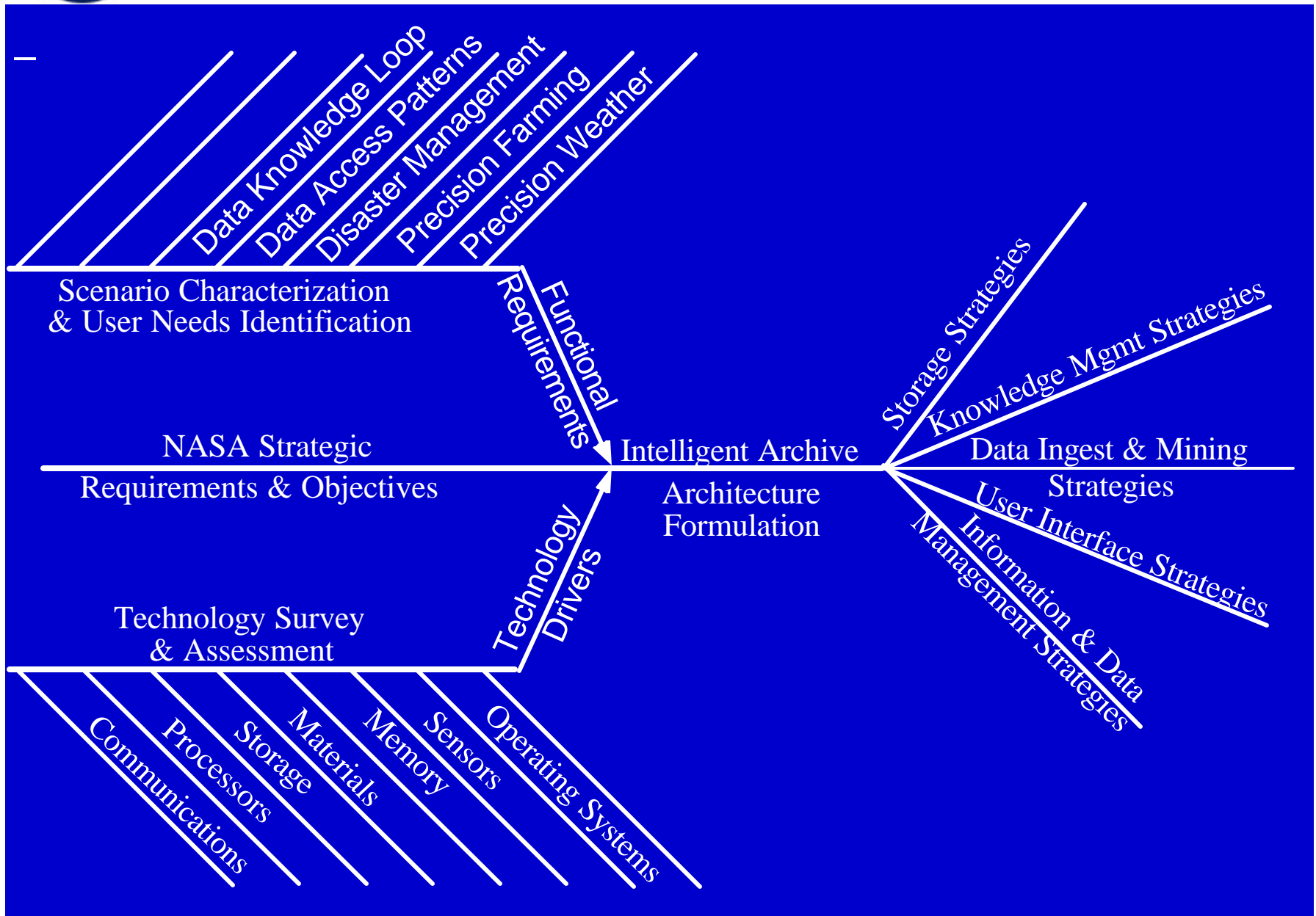
Autonomy in an IA

- Holdings' Management Autonomy
  - Provides data to a science knowledge base in the context of research activities
  - Can exploit and use collected data in the context of a science enterprise
  - Is aware of its data and knowledge holdings and is constantly searching new and existing data for unidentified objects, features or processes
  - Facilitates derivation of information and knowledge using algorithms for Intelligent Data Understanding
  - Works autonomously to identify and characterize objects and events, thus enriching the collections of data, information and knowledge
- User Services Autonomy
  - Recognizes the value of its results, indexes/formats them properly, and delivers them to concerned individuals
  - Interacts with users in human language and visual imagery that can be easily understood by both people and machines
- System Management Autonomy
  - Works with other autonomous information system functions to support research
  - Manages its activities and functions from sensor to user
  - Manages the optimization of its own configuration
  - Observes its own operation and improves its own performance
  - Has awareness of the "state" of its cooperating external partners

Back-up Charts

Approach

Intelligent Archive Architecture Formulation (fishbone diagram)

Upper-left branch — **Scenario Characterization & User Needs Identification** with sub-branches:
- Data Knowledge Loop
- Data Access Patterns
- Disaster Management
- Precision Farming
- Precision Weather

Left spine input: **NASA Strategic Requirements & Objectives**

Lower-left branch — **Technology Survey & Assessment** with sub-branches:
- Communications
- Processors
- Storage
- Materials
- Memory
- Sensors
- Operating Systems

Central connectors:
- Functional Requirements
- Technology Drivers

Right side — **Data Ingest & Mining Strategies** with branches:
- Storage Strategies
- Knowledge Mgmt Strategies
- User Interface Strategies
- Information & Data Management Strategies

NASA Relevance

- Earth Science has large archive holdings that are growing at an ever increasing rate
  - EOSDIS archive just exceeded one petabyte in February of this year
  - New missions (e.g., Aqua) will put additional strains on existing archive services or require additional services
  - User interfacing and data selection are a challenge due to increasing volumes of data and the distributed nature of archives
- Space Science's virtual observatory archiving is expected to be as demanding as Earth science's in the near future
  - Virtual observatory's data volumes will match Earth science's as the program matures
  - Data sources and archives will be distributed (expected to be located close to land based observing instruments)
- The Intelligent Archive Project is formulating strategies and architectures to help resolve the challenges in archiving for Earth and space sciences that result from
  - Ever increasing amount of data volumes and rates
  - Increasing numbers of missions and data sources
  - Increasing demand to support greater numbers of scientists and areas of research
  - Heterogeneous and distributed environment of data providers/users
  - Complexity of data, information and knowledge
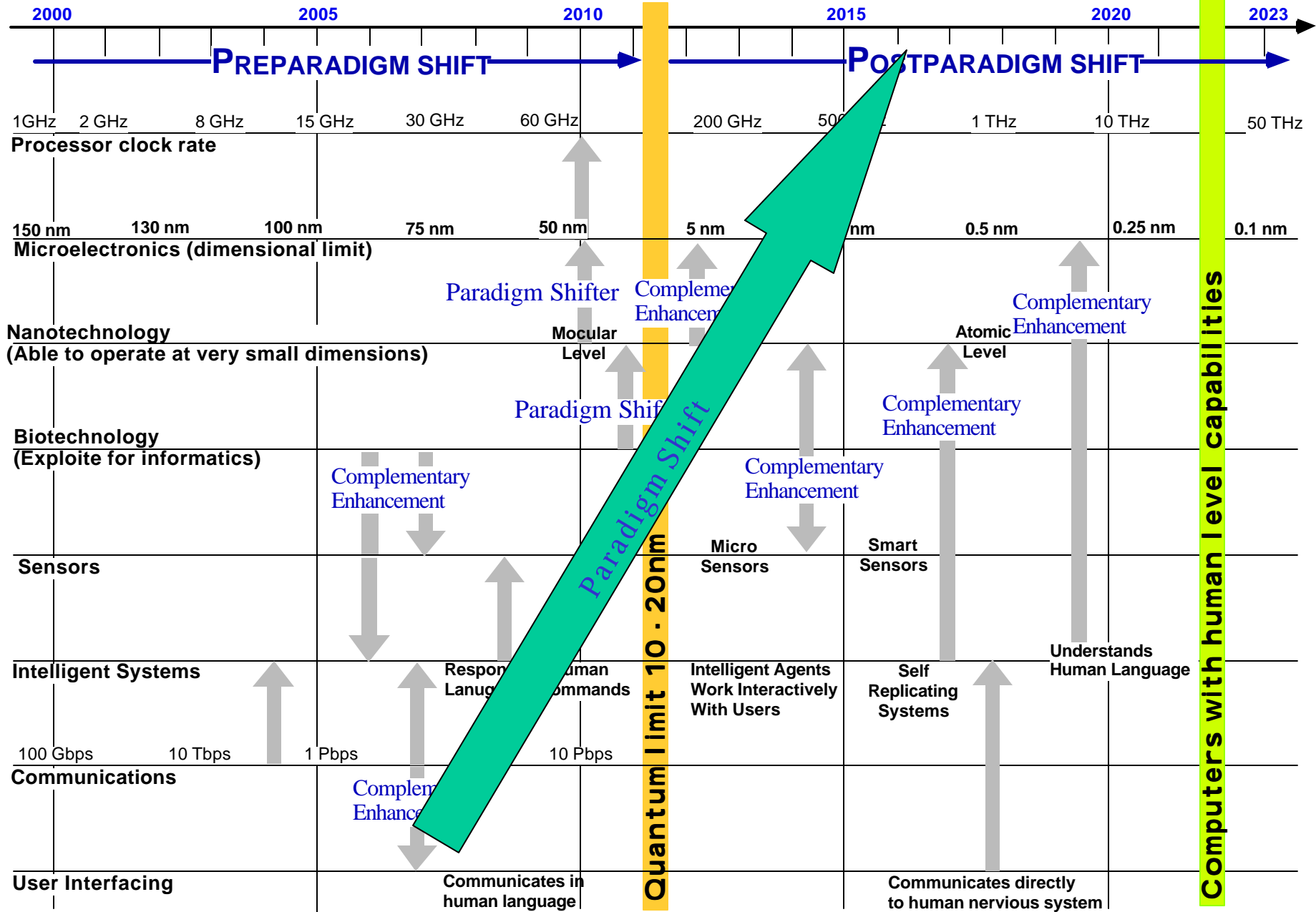
Preliminary Findings

- Future missions will significantly increase the amount and kind of data to be archived and managed

- Expect large numbers of distributed users with personal computing and storage resources that could become a source of data to be archived

- Functions that will need automation and infusion of intelligence
  - Data acquisition, cataloging and characterization
  - Production operations
  - Transformation of data into information
  - User (human and computer) access and communication
  - Forecast and prediction model support
  - Storage and supporting management strategies
  - System management, communications and planning

- Science missions will commonly include models and simulations

- Modeling systems may become an "archive user" that will task sensors, in near real time, to collect data to support simulation analysis,
  - Models could request sensors for specific acquisitions of data
  - Requested data will need to be processed in a timely manner
  - The number of sensors that could be tasked may be large in number, and distributed
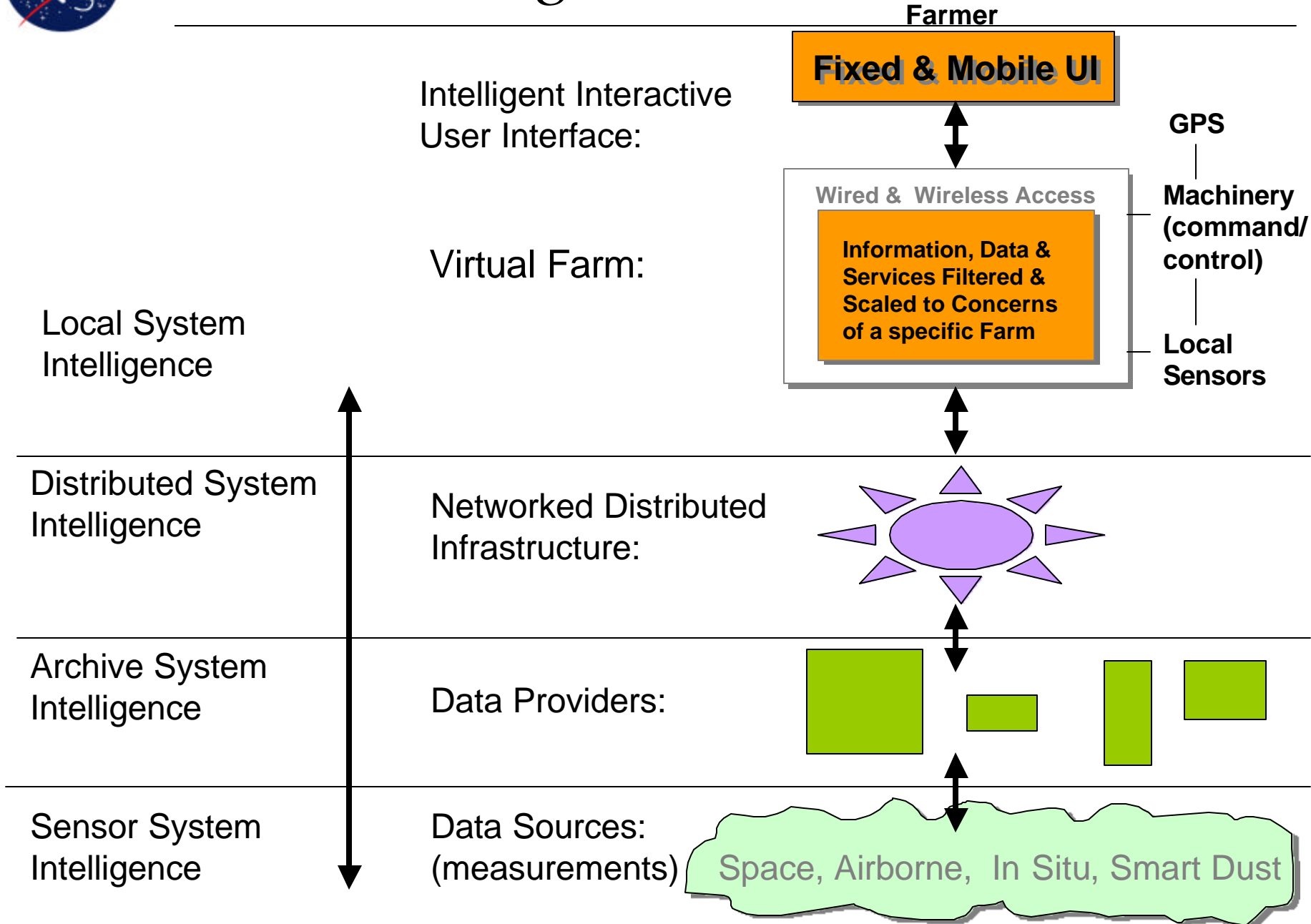
Preliminary Findings (Cont.)

- Use cases have provided valuable information related to archive functionality, scope, and performance
  - Two use cases have been evaluated; precision farming and weather
- Assessed and characterized technology evolution over next 20 years
  - Expect a major paradigm shift (See diagram next viewgraph) which will have a fundamental impact on functionality and performance
- Formulated architectures that relate technologies to core functions
- Topology and texture of architecture very likely to be adaptable and evolutionary
- Archives may store only limited levels of data and produce virtual data products on-demand
- Existing archive systems will need to be integrated into a future intelligent archive
- Intelligent archive will utilize distributed computing to extent possible

Technology Forecast Timeline

Precision Agriculture Scenario

Farmer

Intelligent Interactive User Interface:

Fixed & Mobile UI

GPS

Virtual Farm:

Wired & Wireless Access

Information, Data & Services Filtered & Scaled to Concerns of a specific Farm

Machinery (command/ control)

Local System Intelligence

Local Sensors

Distributed System Intelligence

Networked Distributed Infrastructure:

Archive System Intelligence

Data Providers:

Sensor System Intelligence

Data Sources: (measurements)

Space, Airborne, In Situ, Smart Dust

Data Volumes

- Estimated 1.5 TB per year for a 1000-acre farm, (including satellite and airborne remote sensing data, in situ data, visualizations, modeling, etc.)
- As of 2001 there were 2,158,000 farms in the U.S. (averaging 436 acres each)*
- Total acreage, 9.4×10^8. Result in data stream of 1.4 EB per year for all U.S. farms.
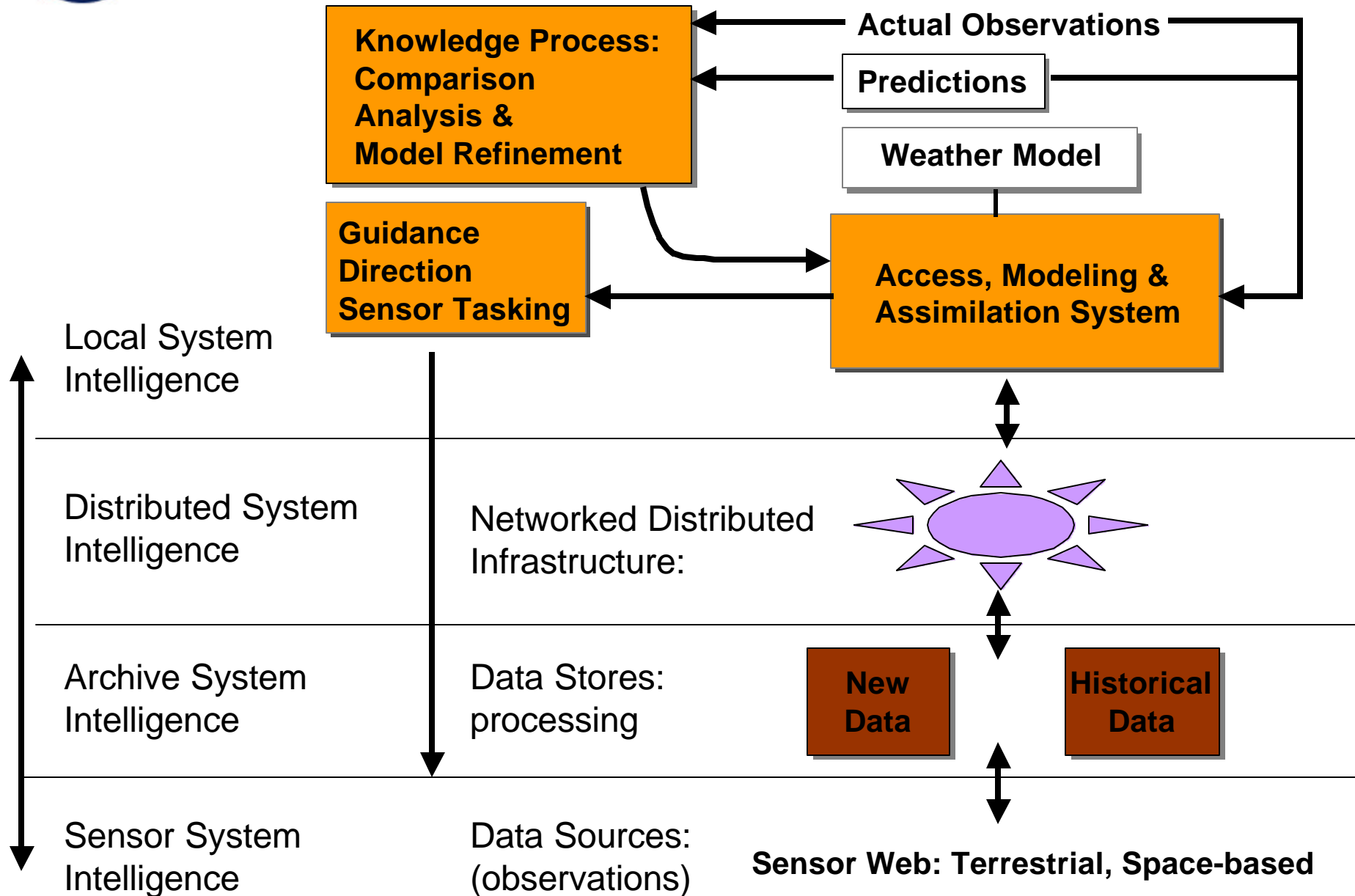
Data Sources

- Remote sensing at spatial resolutions as small as 1 foot and temporal resolutions as fine as 1 hour;
- Precision weather forecasts from three hours ahead to longer-range climate predictions for months and years;
- Land profiles including soils, moisture, elevations, drainage patterns and water sources, digital ortho-photo quadrangles (images with integrated USGS topography maps), digital elevation models, geo-rectified spatial data, ecological zone profiles, biodiversity inventory, local calibrations including ground truth.

Types of analysis needed

- Planting conditions indicators, crop monitoring (growth, maturity, health and stress indicators),
- Weed and pest identification and tracking, chemical and other intervention impact prediction and analysis,
- Weather conditions, advance forecasts, environmental alerts, microclimate surveys, soil types and depths.

* Source: Wisconsin Agricultural Statistics Service via USDA

Precision Weather Scenario

Knowledge Process:
Comparison
Analysis &
Model Refinement

Actual Observations

Predictions

Weather Model

Guidance
Direction
Sensor Tasking

Access, Modeling &
Assimilation System

Local System
Intelligence

Distributed System
Intelligence

Networked Distributed
Infrastructure:

Archive System
Intelligence

Data Stores:
processing

New
Data

Historical
Data

Sensor System
Intelligence

Data Sources:
(observations)

Sensor Web: Terrestrial, Space-based

Precision Weather Support Information

Data Volumes

- Approximately 7.5 TB/day by 2025*

Data Sources

- Space-based, airborne and terrestrial sensors networked via fiber and wireless,
- Large forecast models (forecast models will be approaching 100 million unique grid points with a million observations

Unique Features

- Modeling systems directly task sensors
- Large number of sensors
- Archive is an integral part of the science/modeling process

Types of analysis required

- Structure information in the free atmosphere every 3 hours, every 25 km globally, and vertically from the surface to 80 km altitude
- Global 3D distribution of cloud height, cloud depth, aerosols, water/ice, and suspended precipitation rates
- Land and sea surface temperature, land surface moisture, albedo, vegetation type
- Planetary boundary layer depth

Technical Issues &Risks

Mission Drivers
- The success of future science missions will become increasingly dependent on data archiving services and how well information can be automatically extracted from data
- Future missions can be expected to use large numbers of sensors which will significantly increase the amount and kind of data to be archived and managed
- Data volumes continue to increase rapidly

Automated Analysis and Understanding
- Data mining is not yet able to function at human levels of performance for the identification and classification of features and phenomena
- Machine learning is only marginally successful in acquiring information and knowledge from humans
- Ability of IDU data mining algorithms to perform better than human experts remains untested
- Commercial market tools are limited in handling complex science data

Technical Issues &Risks (Cont.)

Archiving Technologies

- Changes in storage technologies will continue to force constant refreshing of data in archives
  - As the data volumes increase, this will become a very costly long-term processing issue
- Retiring aged systems will be a problem (they have mortality)

- Full automation of archiving function is difficult to achieve
  - Science data complexity often forces significant manual intervention.
  - Standardization of data models would help, but move is toward increasing heterogeneity since populating data models is burdensome
  - Many failures stem from software errors, which are resistant to automated fail-over